IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Egarat K.
September 12, 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

Methodologies

As with any data science projects, the first step was to find and collect the data. Then, we performed data wrangling to get a good overview of the data as a whole and removing unnecessary rows. After that, we dug deeper into the data with Exploratory Data Analysis to identify features and labels. After that we normalized the data and split them between training and test data before we finally trained the models.

Results

This presentation shows the results of methodologies mentioned above and the accuracy of the models that have been trained to classify the data.

Introduction

- SpaceX launches Falcon 9 rockets at a cost of around $62m. This is considerably cheaper than other providers (which usually cost upwards of $165m), and much of the savings are because SpaceX can land, and then re-use the first stage of the rocket.
- If we can make predictions on whether the first stage will land, we can determine the cost of a launch, and use this information to assess whether an alternate company should bid and SpaceX for a rocket launch.
- This project will ultimately predict if the Space X Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

Data collection methodology:

- Collecting the data from the SpaceX REST API
- Scraping rocket launch data from Wikipedia

Perform data wrangling

- Remove NaN and Null values from the dataset
- Determine number of launches, orbit, mission outcome to help identify the weight of these data
- Create an outcome label on historical data that will be predicted by the model for future data

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

- Use scikit-learn to preprocess, split the data for training and testing, train classification models
- Plot confusion matrices and assessing the accuracy of them

Data Collection

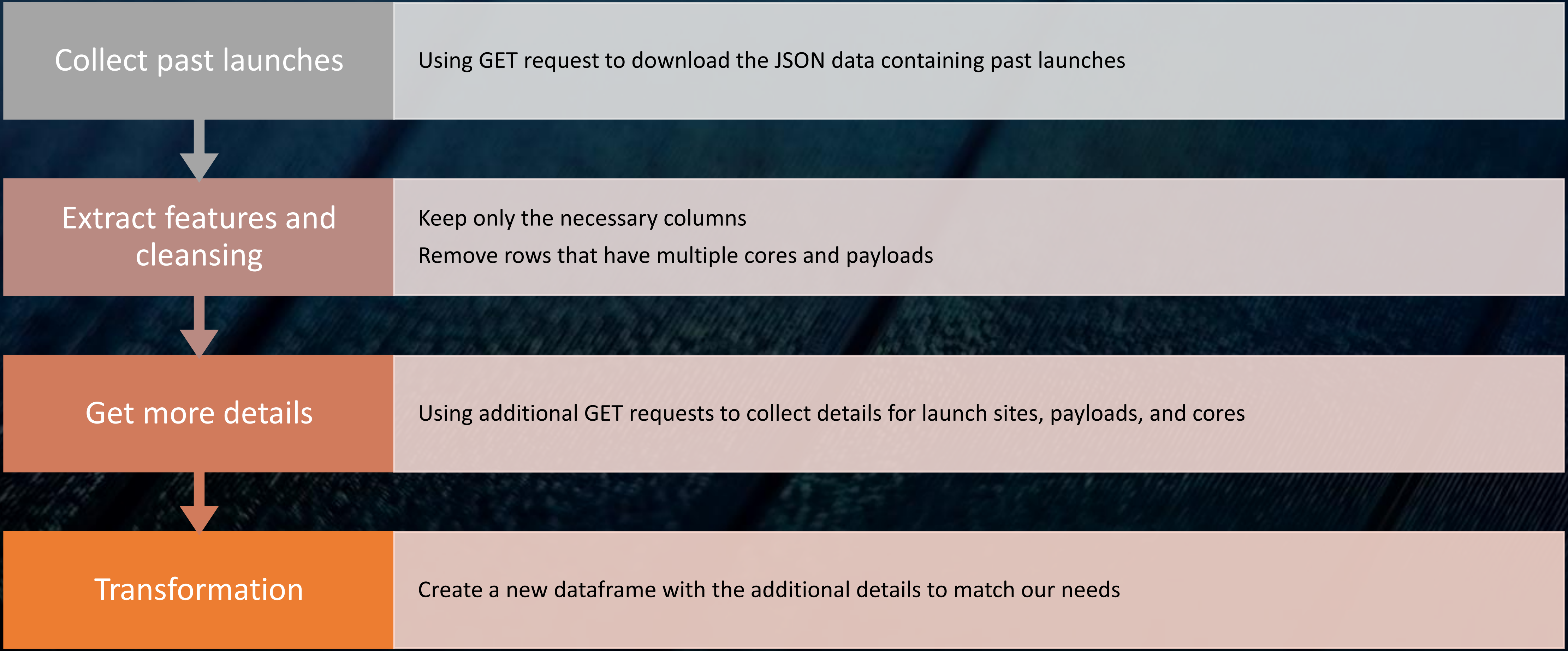
The following slides will show the process on how the data has been collected and prepared to support further data science activities.

The data were gathered from the following sources:

- SpaceX REST API
- Wikipedia article for Falcon 9 launches

Clicking on the link will show the related notebooks.

Data Collection – SpaceX API



Collect past launches

Using GET request to download the JSON data containing past launches

Extract features and cleansing

Keep only the necessary columns

Remove rows that have multiple cores and payloads

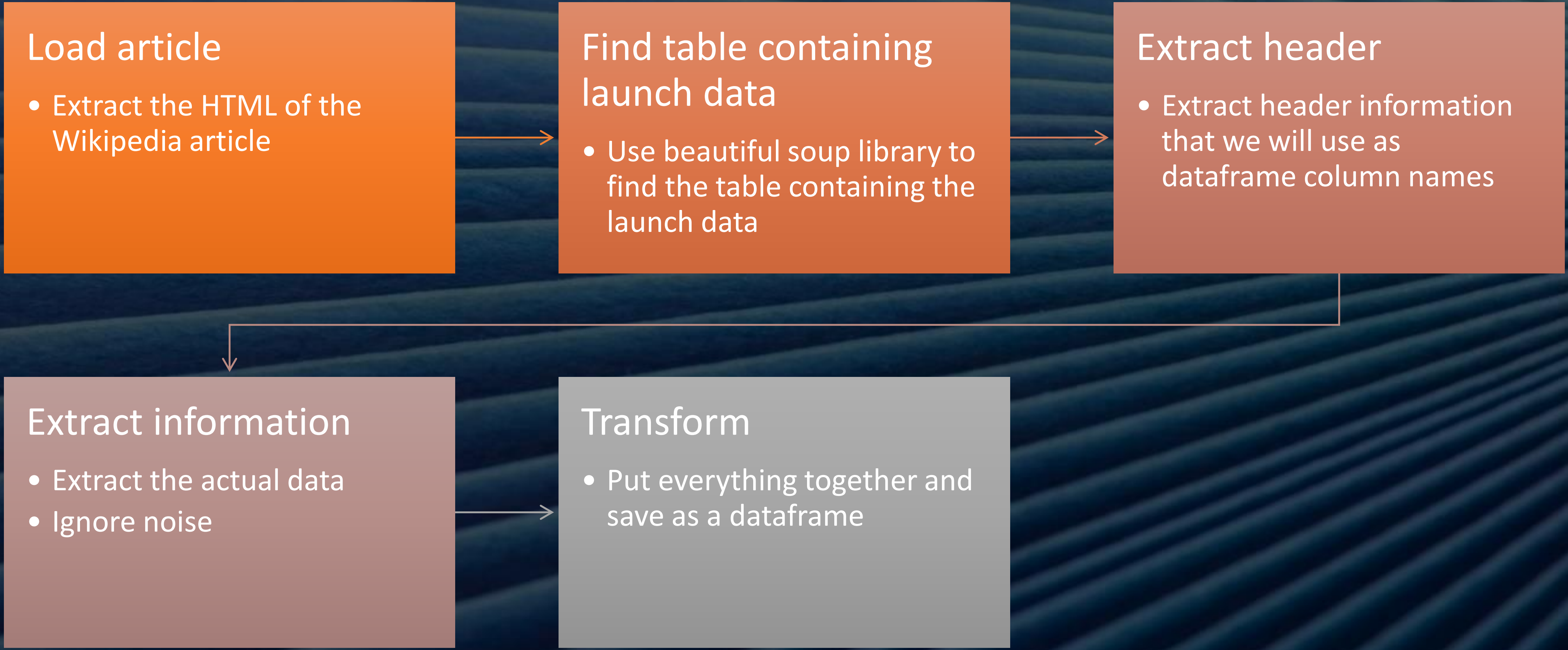
Get more details

Using additional GET requests to collect details for launch sites, payloads, and cores

Transformation

Create a new dataframe with the additional details to match our needs

Data Collection - Scraping



Load article

- Extract the HTML of the Wikipedia article

Find table containing launch data

- Use beautiful soup library to find the table containing the launch data

Extract header

- Extract header information that we will use as dataframe column names

Extract information

- Extract the actual data
- Ignore noise

Transform

- Put everything together and save as a dataframe

Data Wrangling

In the data wrangling process, we found out about the following:

- Missing values in the data
- Number of launches per site
- Number of launches per orbit

Furthermore, we added a column for the classification label in our dataset. The label will be used for model training and is the one that we want to predict in the future.

Data Wrangling Notebook

EDA with Data Visualization

SCATTER CHARTS

Scatter charts were produced to visualize the relationships between:

- Flight Number and Launch Site
- Payload and Launch Site
- Orbit Type and Flight Number
- Payload and Orbit Type

Scatter charts are useful to observe relationships, or correlations, between two numeric variables.

BAR CHART

A bar chart was produced to visualize the relationship between:

- Success Rate and Orbit Type

Bar charts are used to compare a numerical value to a categorical variable. Horizontal or vertical bar charts can be used, depending on the size of the data.

LINE CHARTS

Line charts were produced to visualize the relationships between:

- Success Rate and Year (i.e. the launch success yearly trend)

Line charts contain numerical values on both axes, and are generally used to show the change of a variable over time.

EDA with Data Visualization Notebook

EDA with SQL

The SQL queries performed on the data set were used to:

1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string ‘CCA’
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display the average payload mass carried by booster version F9 v1.1
5. List the date when the first successful landing outcome on a ground pad was achieved
6. List the names of the boosters which had success on a drone ship and a payload mass between 4000 and 6000 kg
7. List the total number of successful and failed mission outcomes
8. List the names of the booster versions which have carried the maximum payload mass
9. List the failed landing outcomes on drone ships, their booster versions, and launch site names for 2015
10. Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

EDA with SQL Notebook

Build an Interactive Map with Folium

The following steps were taken to visualize the launch data on an interactive map:

1. Mark all launch sites on a map
   - Initialise the map using a Folium Map object
   - Add a folium.Circle and folium.Marker for each launch site on the launch map

2. Mark the success/failed launches for each site on a map
   - As many launches have the same coordinates, it makes sense to cluster them together.
   - Before clustering them, assign a marker colour of successful (class = 1) as green, and failed (class = 0) as red.
   - To put the launches into clusters, for each launch, add a folium.Marker to the MarkerCluster() object.
   - Create an icon as a text label, assigning the icon_color as the marker_colour determined previously.

3. Calculate the distances between a launch site to its proximities
   - To explore the proximities of launch sites, calculations of distances between points can be made using the Lat and Long values.
   - After marking a point using the Lat and Long values, create a folium.Marker object to show the distance.
   - To display the distance line between two points, draw a folium.PolyLine and add this to the map.

Interactive Map with Folium Notebook

Build a Dashboard with Plotly Dash

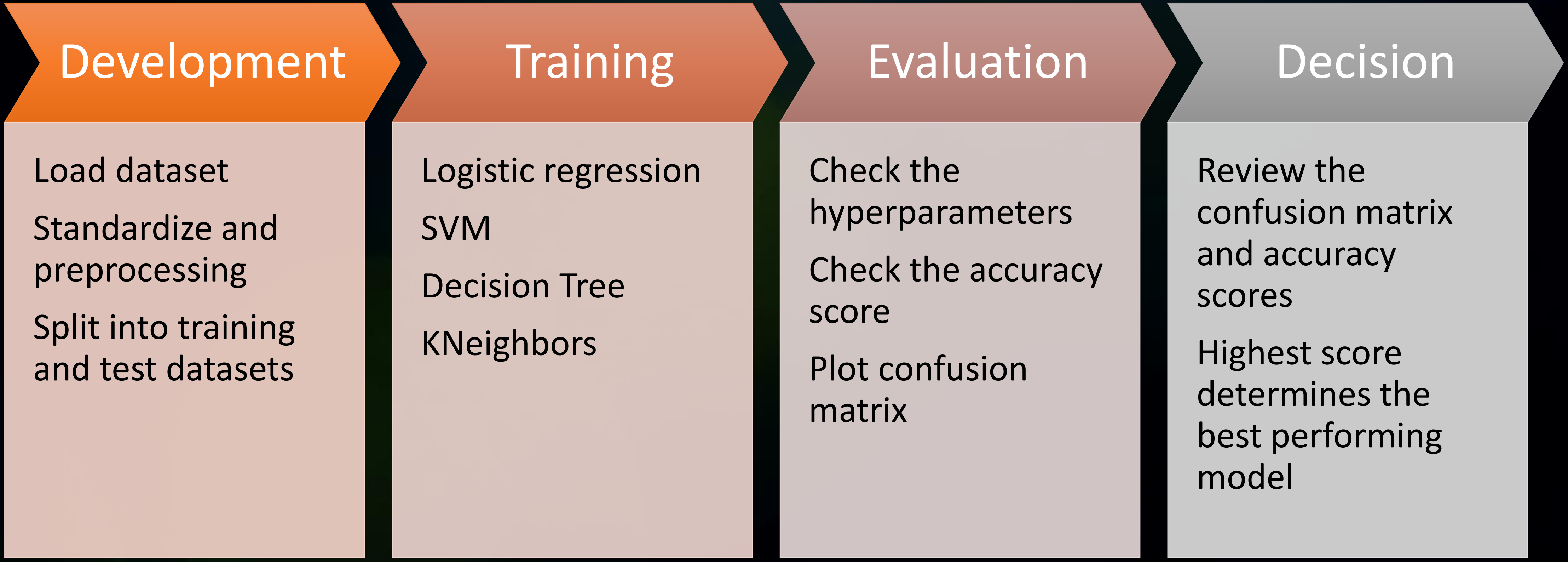
The following plots were added to a Plotly Dash dashboard to have an interactive visualisation of the data:

1. Pie chart (px.pie()) showing the total successful launches per site
   - This makes it clear to see which sites are most successful
   - The chart could also be filtered (using a dcc.Dropdown() object) to see the success/failure ratio for an individual site

2. Scatter graph (px.scatter()) to show the correlation between outcome (success or not) and payload mass (kg)
   - This could be filtered (using a RangeSlider() object) by ranges of payload masses
   - It could also be filtered by booster version

Plotly script

Predictive Analysis (Classification)



Development

- Load dataset
- Standardize and preprocessing
- Split into training and test datasets

Training

- Logistic regression
- SVM
- Decision Tree
- KNeighbors

Evaluation

- Check the hyperparameters
- Check the accuracy score
- Plot confusion matrix

Decision

- Review the confusion matrix and accuracy scores
- Highest score determines the best performing model

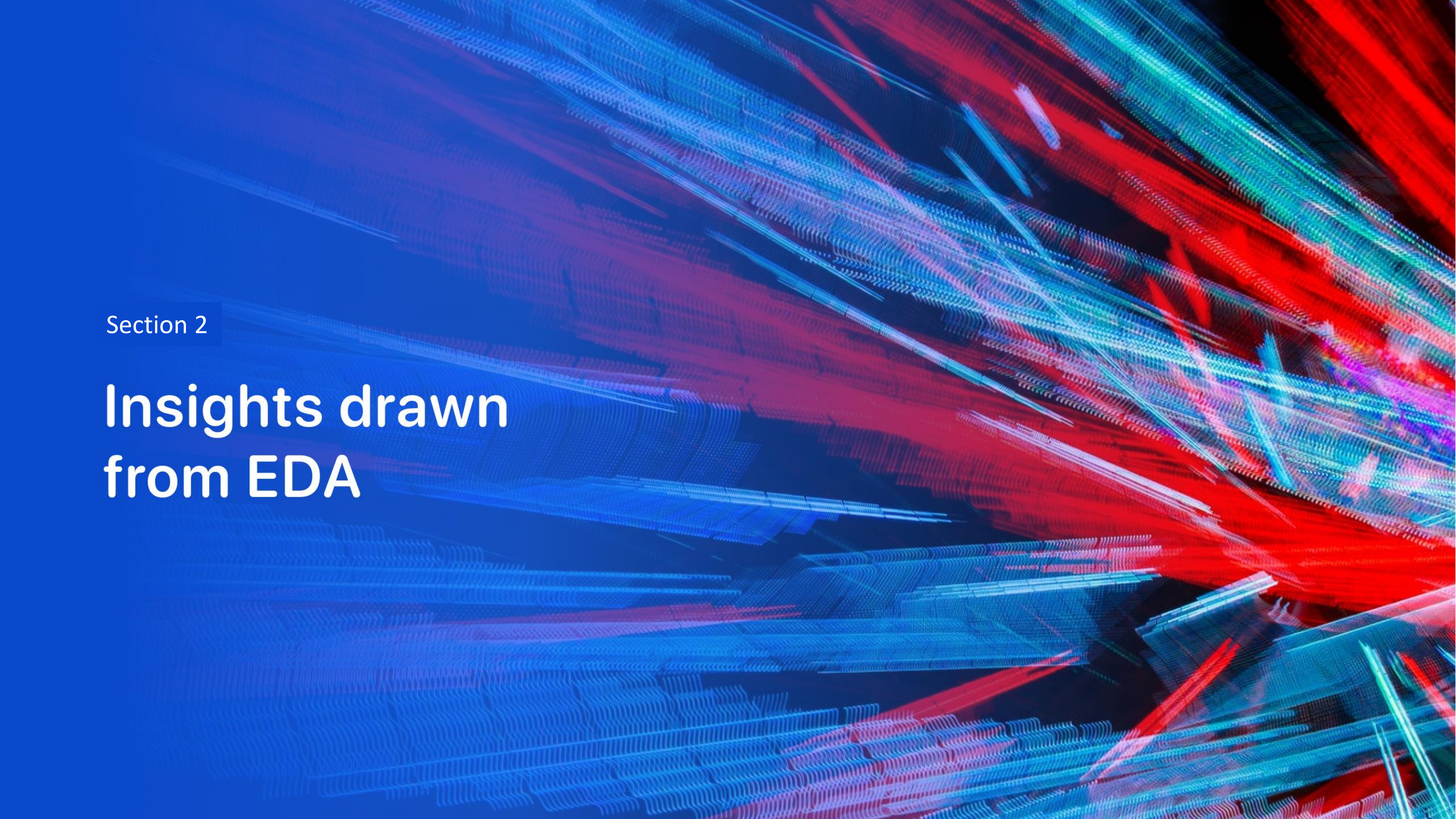
Predictive Analysis Notebook

Results

EXPLORATORY DATA ANALYSIS

INTERACTIVE ANALYTICS

PREDICTIVE ANALYSIS



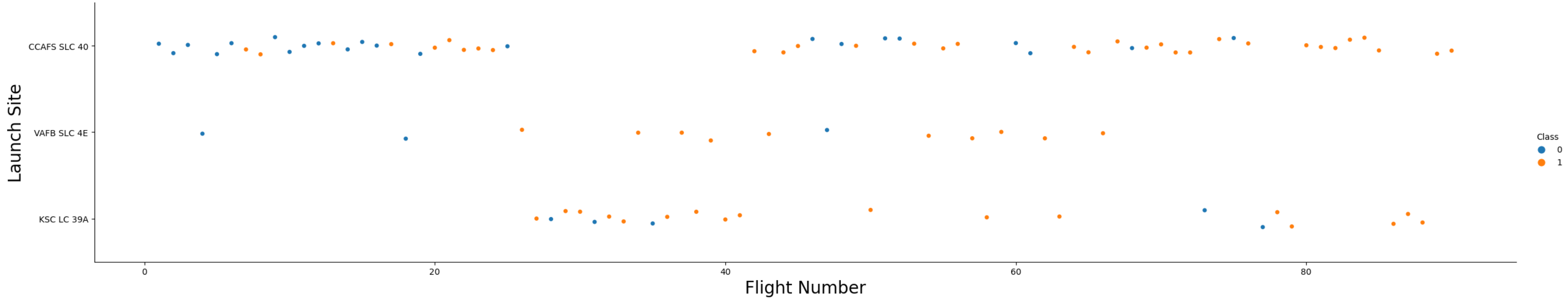
Section 2

Insights drawn from EDA

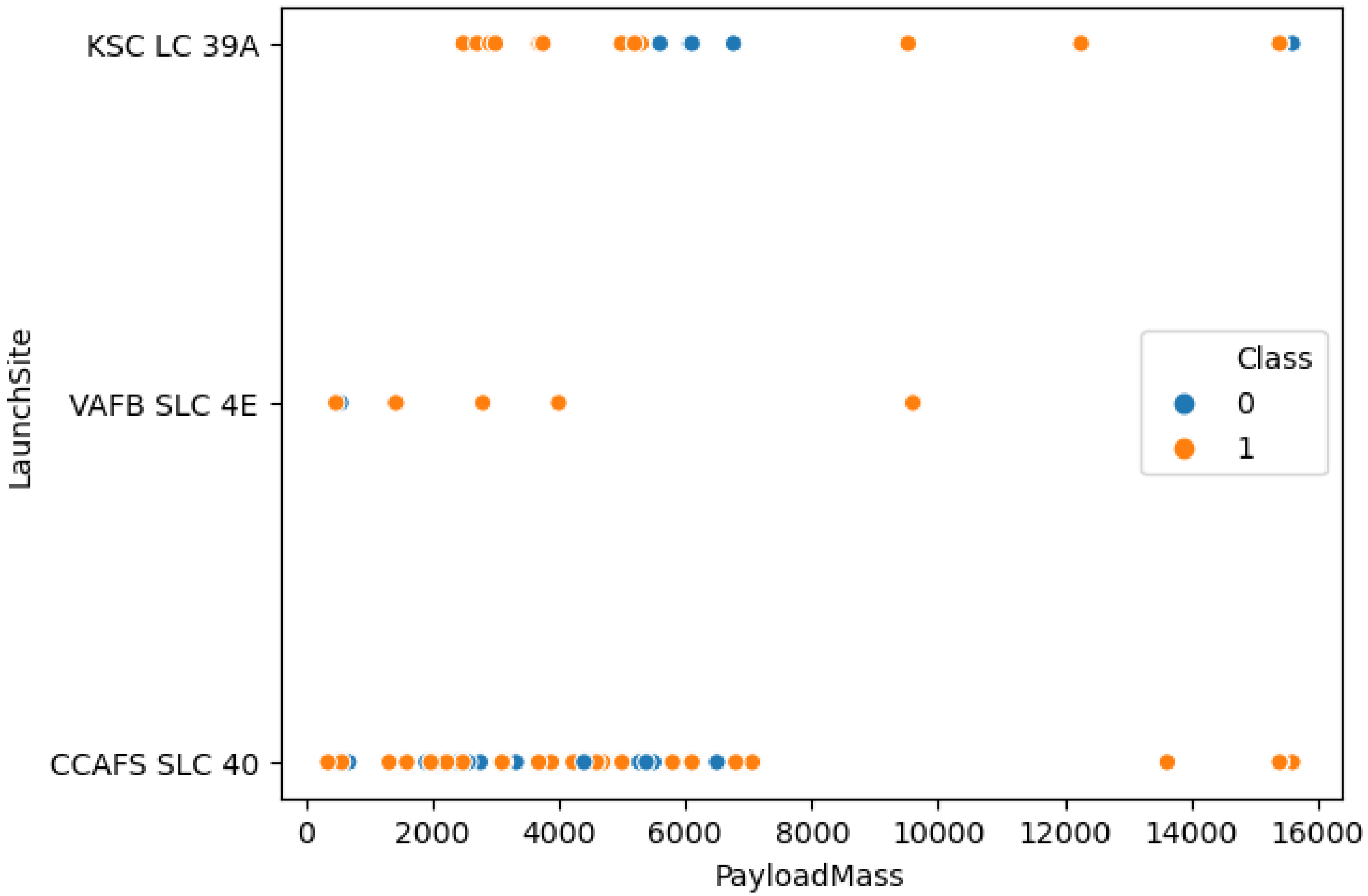
Flight Number vs. Launch Site

The scatter plot of Launch Site vs. Flight Number shows that:

- As the number of flights increases, the rate of success at a launch site increases.
- Most of the early flights (flight numbers < 30) were launched from CCAFS SLC 40, and were generally unsuccessful.
- The flights from VAFB SLC 4E also show this trend, that earlier flights were less successful.
- No early flights were launched from KSC LC 39A, so the launches from this site are more successful.
- Above a flight number of around 30, there are significantly more successful landings (Class = 1).

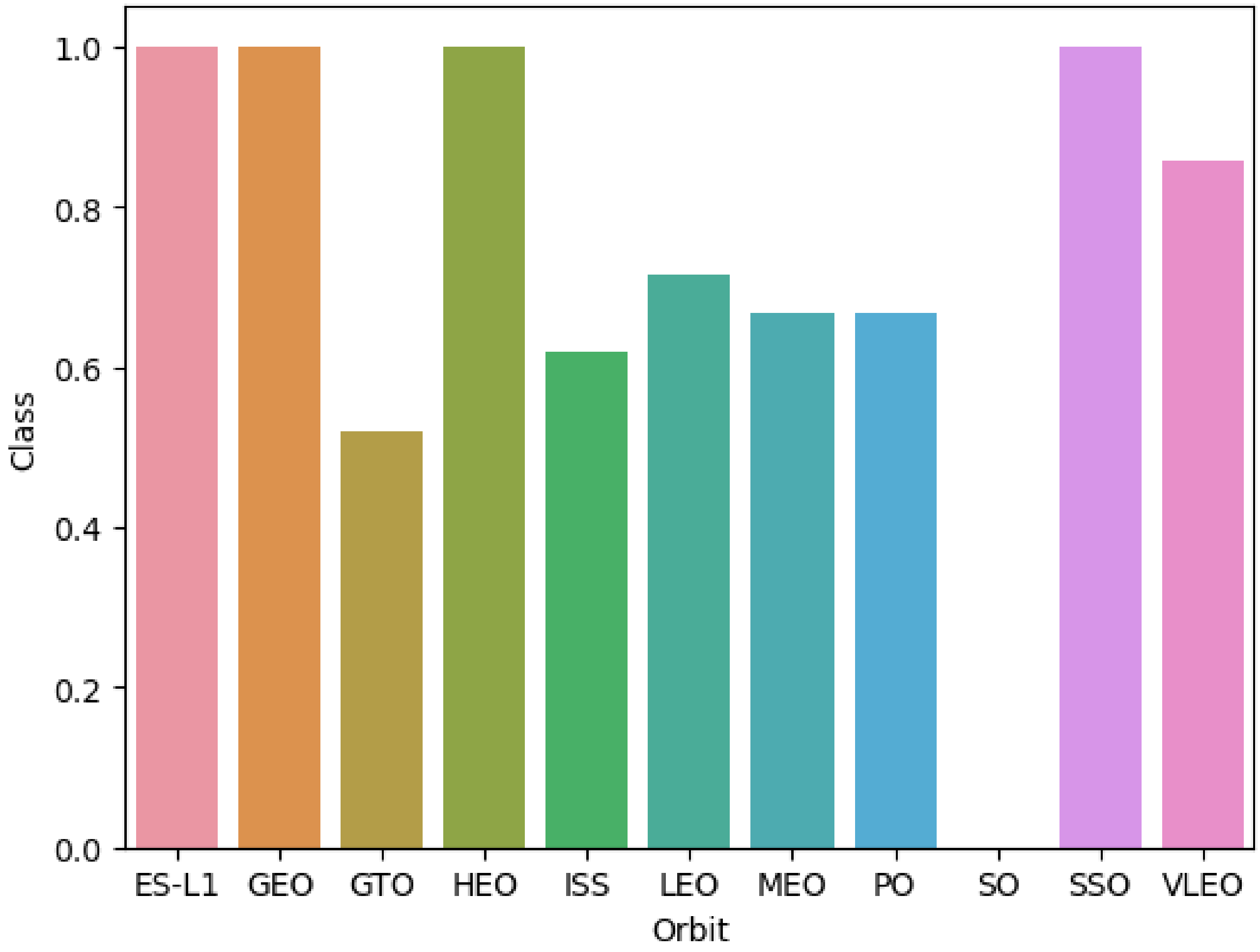


Payload vs. Launch Site



The scatter plot of Launch Site vs. Payload Mass shows that:

- Above a payload mass of around 7000 kg, there are very few unsuccessful landings, but there is also far less data for these heavier launches.
- There is no clear correlation between payload mass and success rate for a given launch site.
- All sites launched a variety of payload masses, with most of the launches from CCAFS SLC 40 being comparatively lighter payloads (with some outliers).



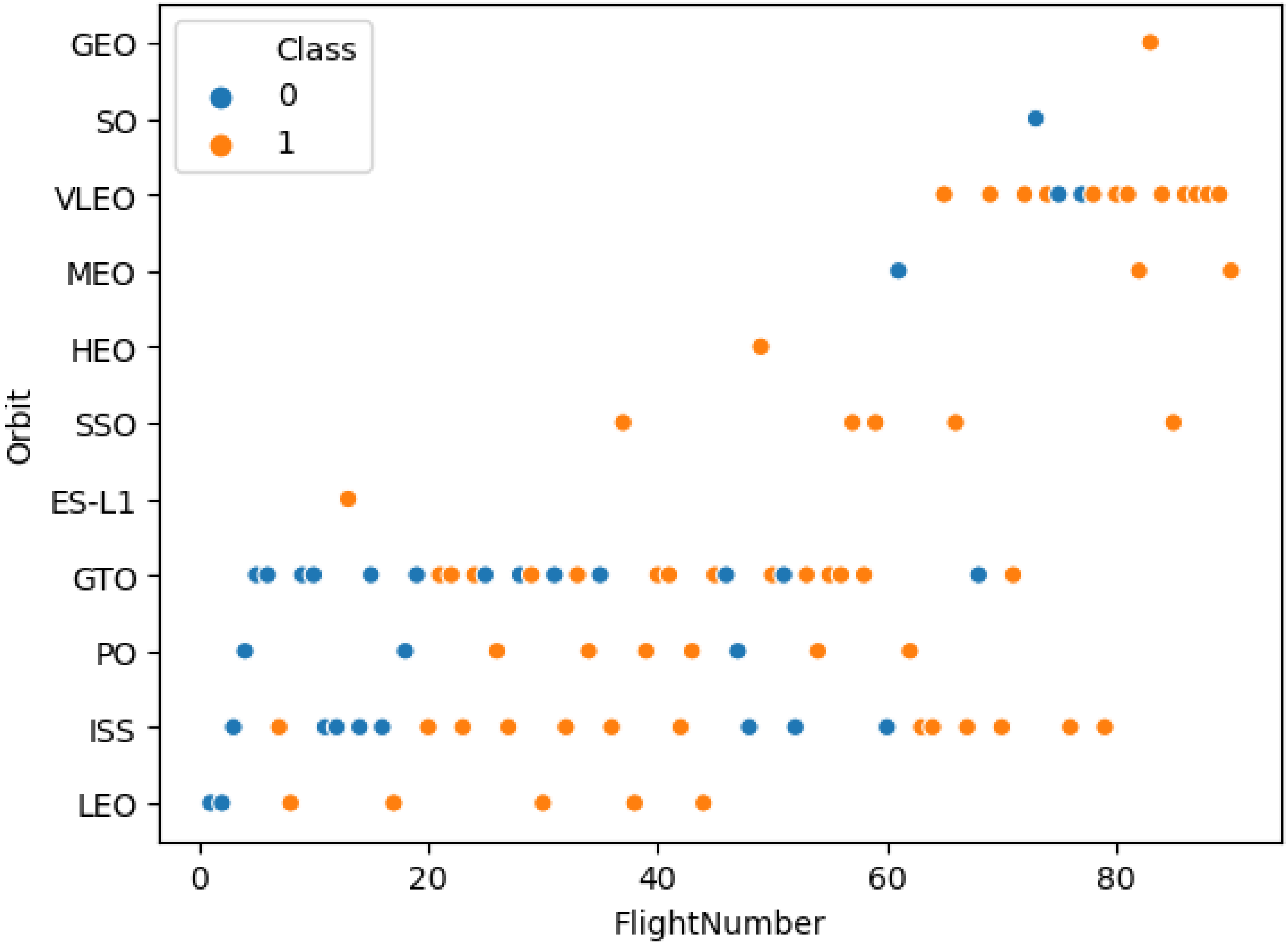
Success Rate vs. Orbit Type

- The bar chart of Success Rate vs. Orbit Type shows that the following orbits have the highest (100%) success rate:
- ES-L1 (Earth-Sun First Lagrangian Point)
- GEO (Geostationary Orbit)
- HEO (High Earth Orbit)
- SSO (Sun-synchronous Orbit)

- The orbit with the lowest (0%) success rate is:
- SO (Heliocentric Orbit)

Flight Number vs. Orbit Type

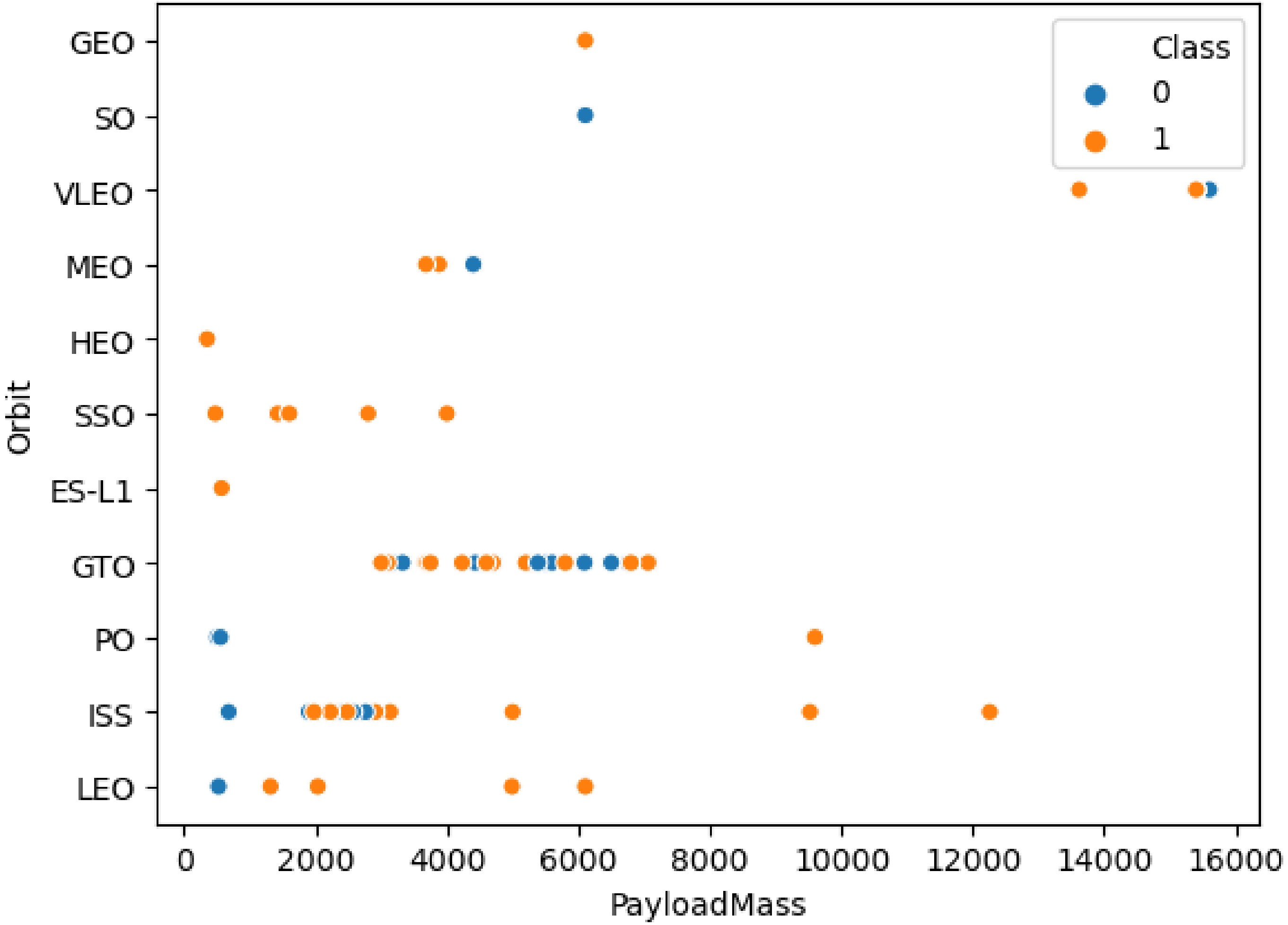
- This scatter plot of Orbit Type vs. Flight number shows a few useful things that the previous plots did not, such as:
  - The 100% success rate of GEO, HEO, and ES-L1 orbits can be explained by only having 1 flight into the respective orbits.
  - The 100% success rate in SSO is more impressive, with 5 successful flights.
  - There is little relationship between Flight Number and Success Rate for GTO.
  - Generally, as Flight Number increases, the success rate increases. This is most extreme for LEO, where unsuccessful landings only occurred for the low flight numbers (early launches).



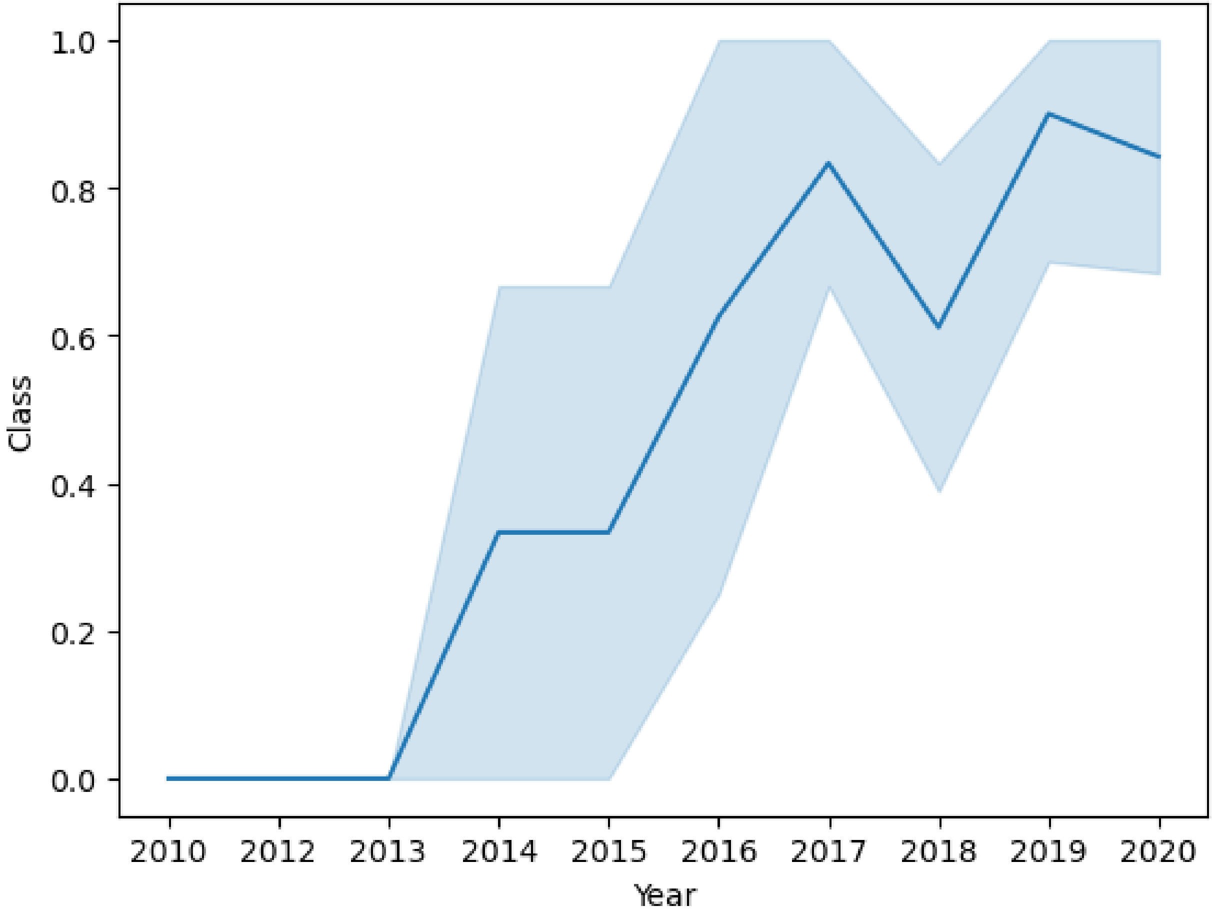
Payload vs. Orbit Type

This scatter plot of Orbit Type vs. Payload Mass shows that:

- The following orbit types have more success with heavy payloads:
  - PO (although the number of data points is small)
  - ISS
  - LEO
- For GTO, the relationship between payload mass and success rate is unclear.
- VLEO (Very Low Earth Orbit) launches are associated with heavier payloads, which makes intuitive sense.



Launch Success Yearly Trend

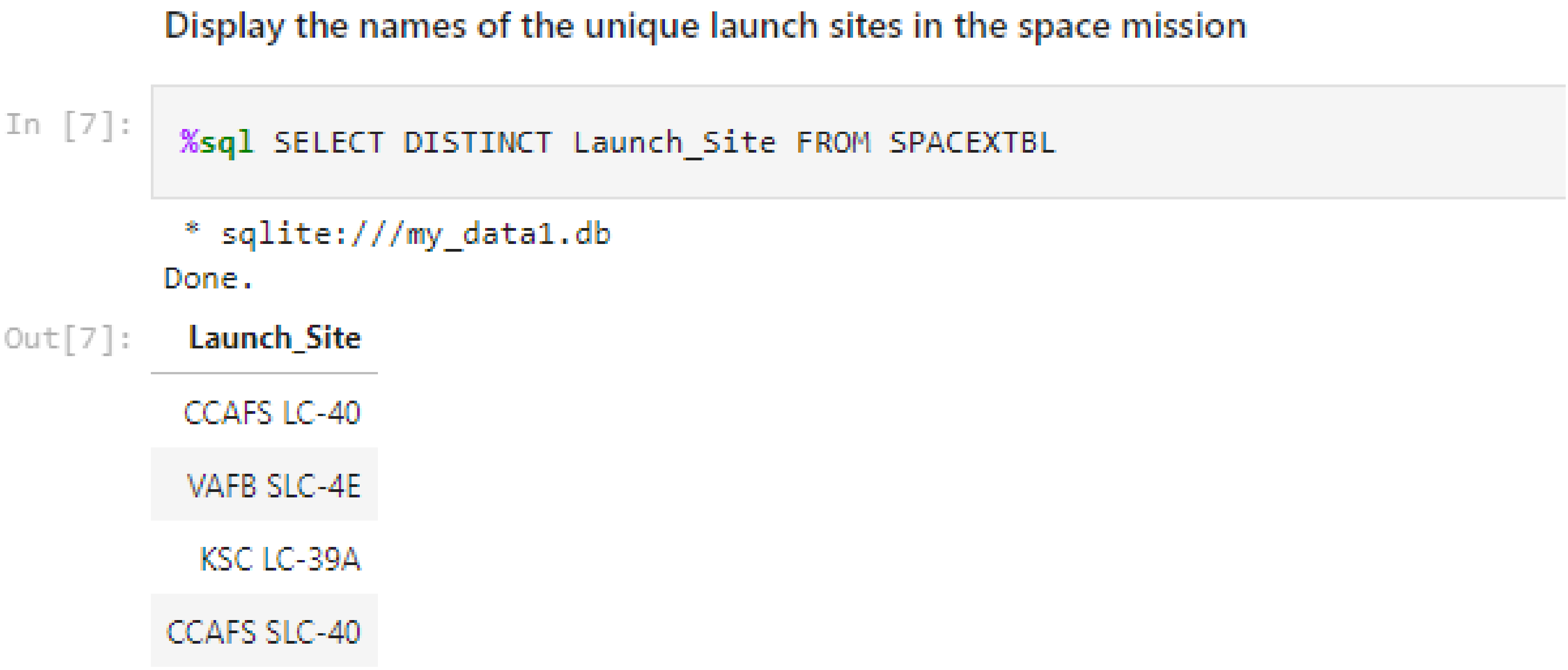


The line chart of yearly average success rate shows that:

- Between 2010 and 2013, all landings were unsuccessful (as the success rate is 0).
- After 2013, the success rate generally increased, despite small dips in 2018 and 2020.
- After 2016, there was always a greater than 50% chance of success.

All Launch Site Names

The keyword DISTINCT shows all unique values in the Launch_Site column which holds the launch site names.



Display the names of the unique launch sites in the space mission

```
In [7]:  %sql SELECT DISTINCT Launch_Site FROM SPACEXTBL
```

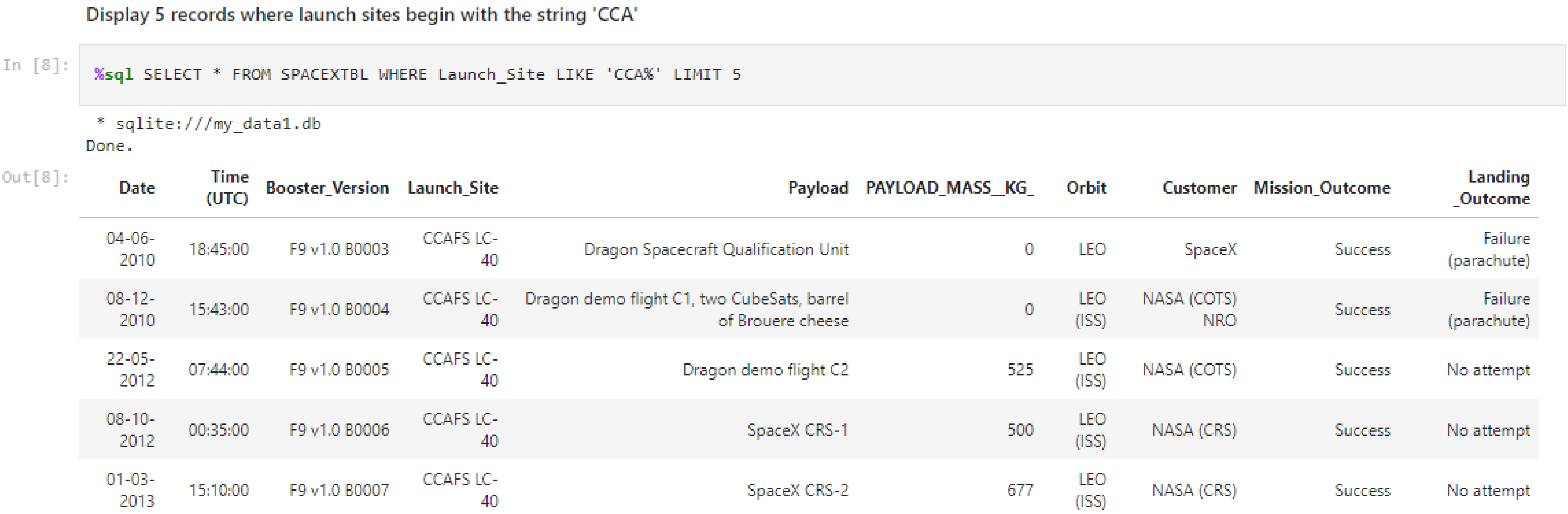
```
 * sqlite:///my_data1.db
Done.
```

Out[7]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- The LIKE ‘CCA%’ predicate acts as a wildcard to search all launch sites starting with CCA
- LIMIT 5 will restrict the output to 5 rows



Display 5 records where launch sites begin with the string 'CCA'

```
In [8]: %sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5
```

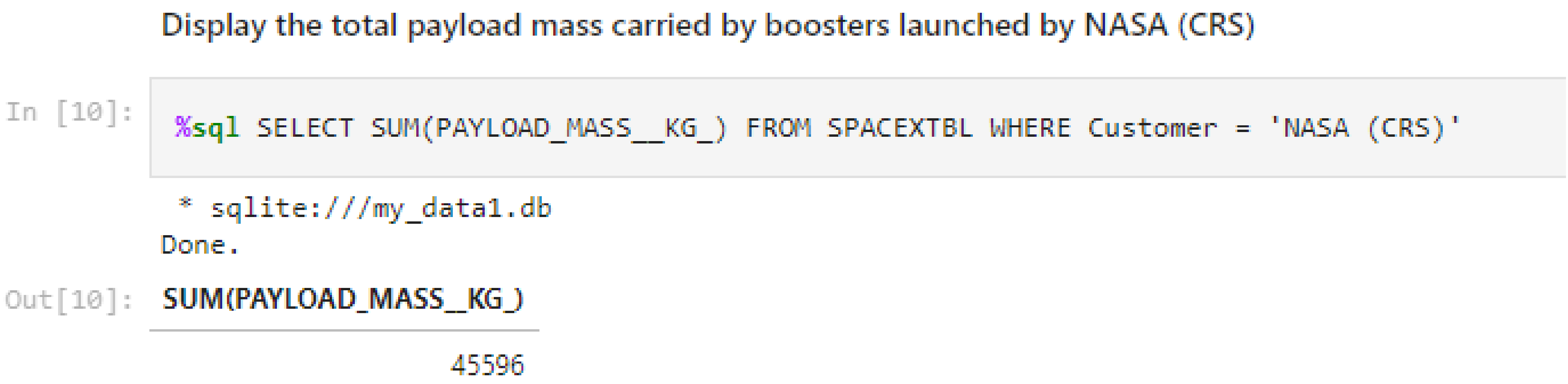
```
 * sqlite:///my_data1.db
Done.
```

Out[8]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The aggregation function SUM() will add the values of the given column
- In the WHERE clause, we have limited to summation of only customers with the value “NASA (CRS)



Display the total payload mass carried by boosters launched by NASA (CRS)

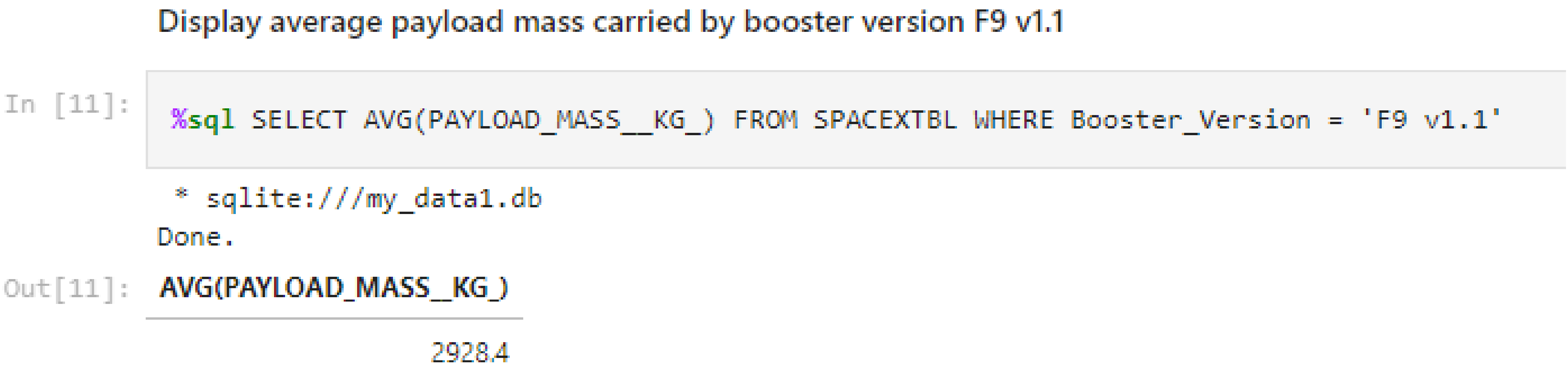
```
In [10]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'

 * sqlite:///my_data1.db
Done.

Out[10]: SUM(PAYLOAD_MASS__KG_)
                          45596
```

Average Payload Mass by F9 v1.1

- Same principle as in the previous slide
- The only difference is that we use the AVG() aggregation function which calculates the mean of the provided column



Display average payload mass carried by booster version F9 v1.1

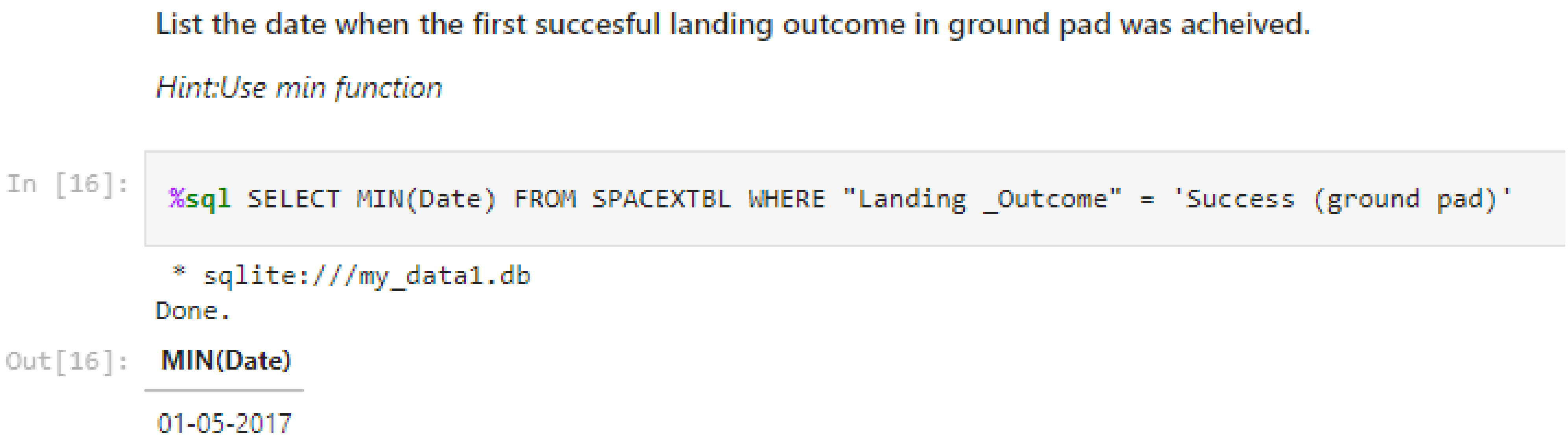
```
In [11]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version = 'F9 v1.1'

 * sqlite:///my_data1.db
Done.

Out[11]: AVG(PAYLOAD_MASS__KG_)
                        2928.4
```

First Successful Ground Landing Date

- The MIN() function returns the earliest date of a row with landing outcome “Success (ground pad)”



List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
In [16]:  %sql SELECT MIN(Date) FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (ground pad)'
```

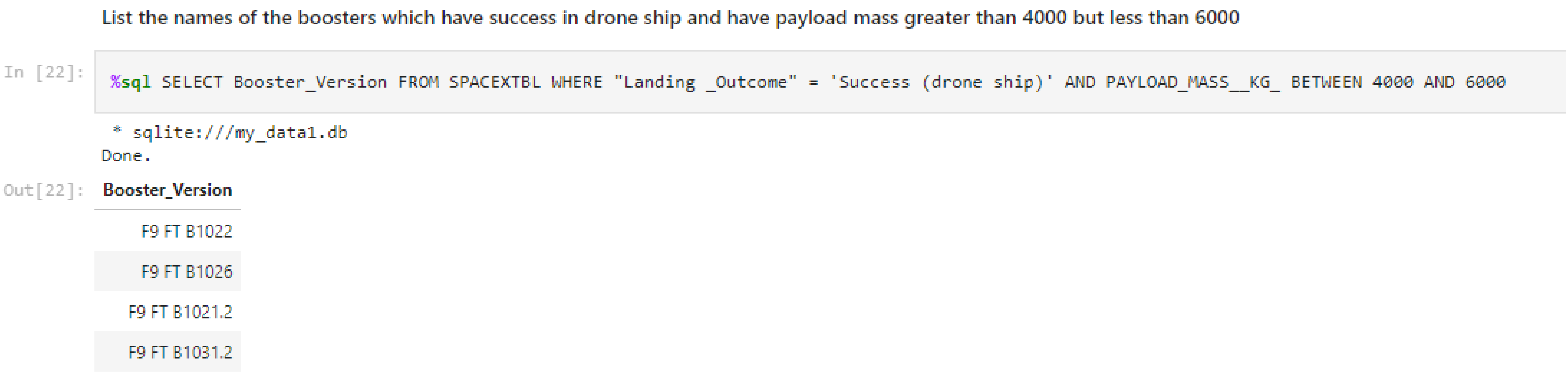
```
 * sqlite:///my_data1.db
Done.
```

Out[16]:

MIN(Date)
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

- The BETWEEN keyword selects all the payload mass greater than 4000 and less than 6000



List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
In [22]: %sql SELECT Booster_Version FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

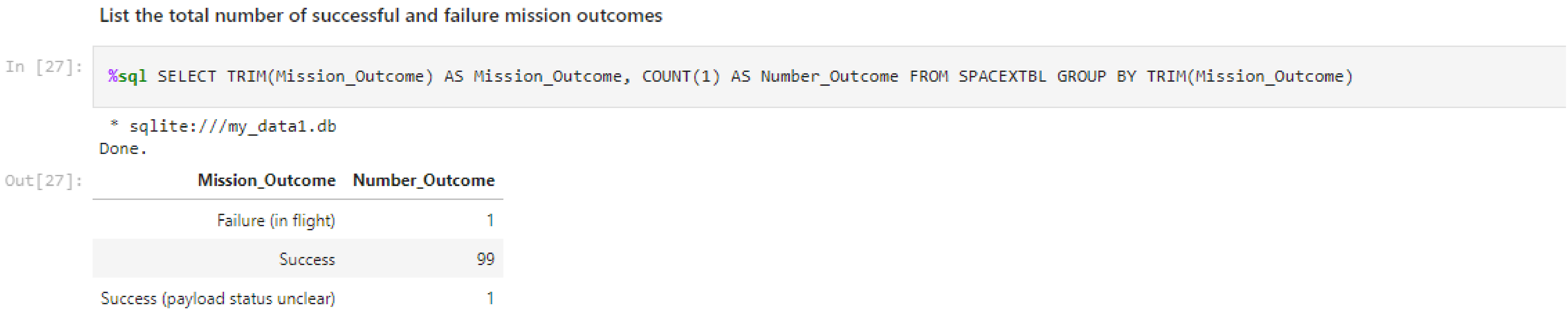
```
 * sqlite:///my_data1.db
Done.
```

Out[22]:

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- The aggregation function COUNT() counts every occurrences of the given criteria
- The GROUP BY keyword groups the result by its mission outcome



List the total number of successful and failure mission outcomes

```
In [27]: %sql SELECT TRIM(Mission_Outcome) AS Mission_Outcome, COUNT(1) AS Number_Outcome FROM SPACEXTBL GROUP BY TRIM(Mission_Outcome)
```

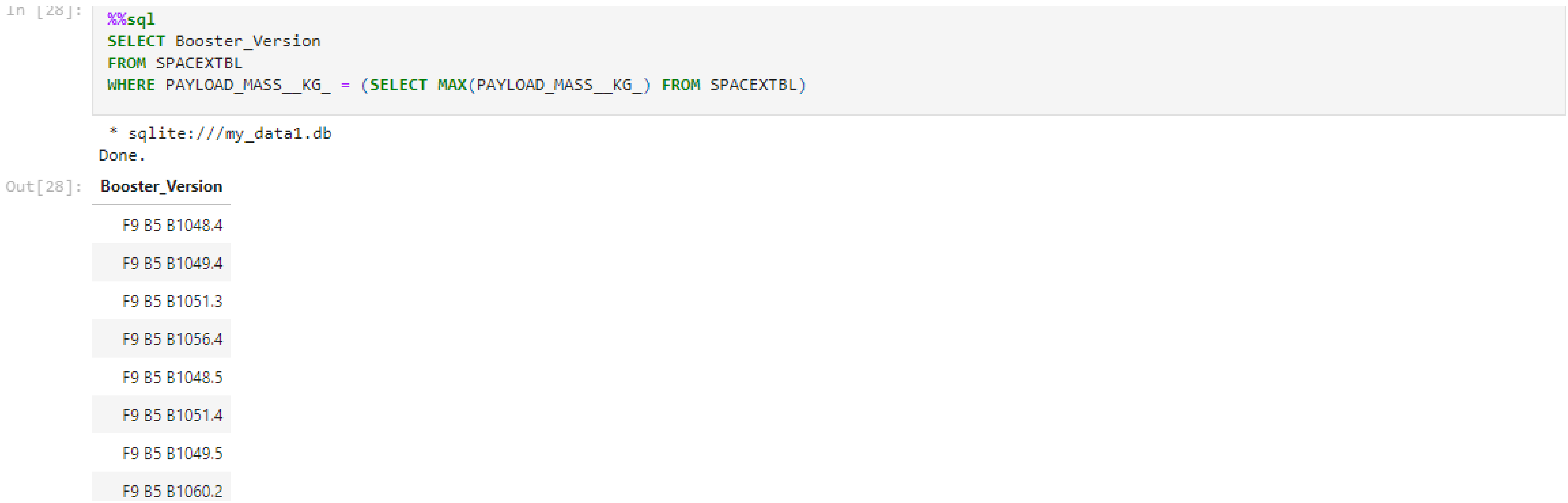
```
 * sqlite:///my_data1.db
Done.
```

Out[27]:

Mission_Outcome	Number_Outcome
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The subquery returns the maximum payload mass with the aggregation function MAX()
- The result of the subquery is used as a criteria for the main query



In [28]:

```
%%sql
SELECT Booster_Version
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

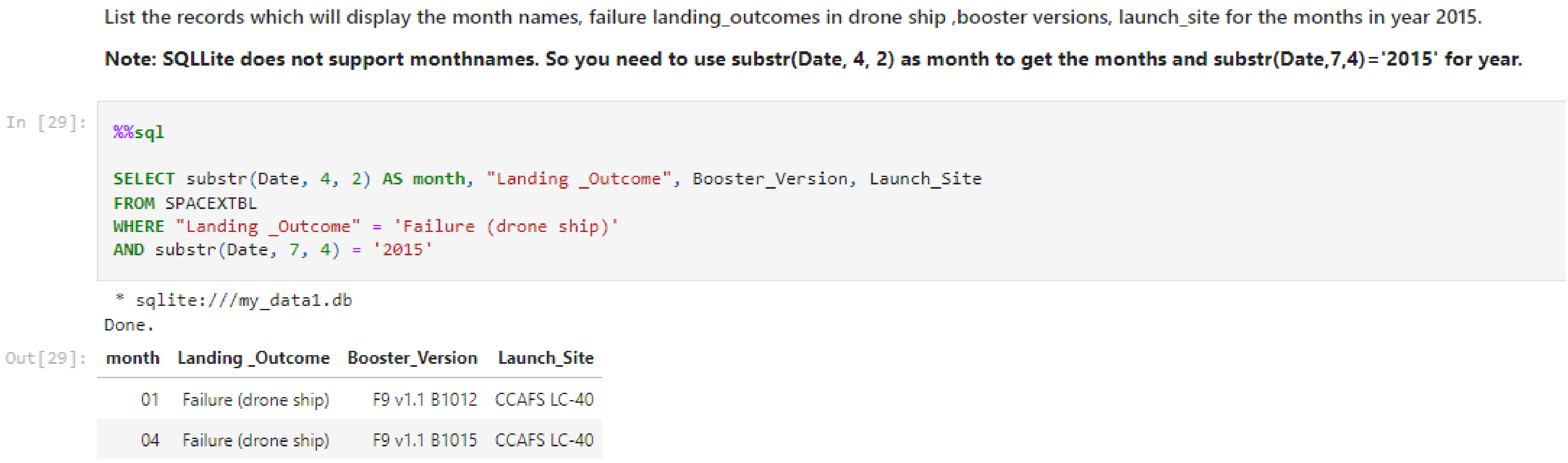
```
 * sqlite:///my_data1.db
Done.
```

Out[28]:

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2

2015 Launch Records

- SQLLite does not have date extraction functions built-in
- To extract the year and month of a date, a substring needs to be selected of the date



List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

In [29]:

```
%%sql

SELECT substr(Date, 4, 2) AS month, "Landing _Outcome", Booster_Version, Launch_Site
FROM SPACEXTBL
WHERE "Landing _Outcome" = 'Failure (drone ship)'
AND substr(Date, 7, 4) = '2015'
```

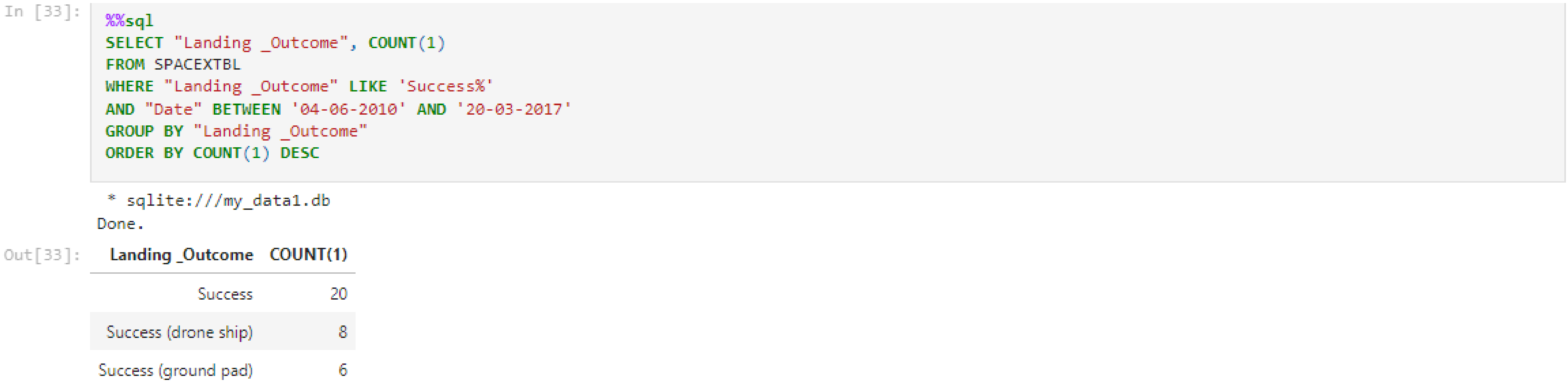
```
 * sqlite:///my_data1.db
Done.
```

Out[29]:

month	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- All concepts are applied in this query
- Worth mentioning is the ORDER BY keyword that is used to sort the result by the COUNT in descending order



In [33]:

```
%%sql
SELECT "Landing _Outcome", COUNT(1)
FROM SPACEXTBL
WHERE "Landing _Outcome" LIKE 'Success%'
AND "Date" BETWEEN '04-06-2010' AND '20-03-2017'
GROUP BY "Landing _Outcome"
ORDER BY COUNT(1) DESC
```

```
 * sqlite:///my_data1.db
Done.
```

Out[33]:

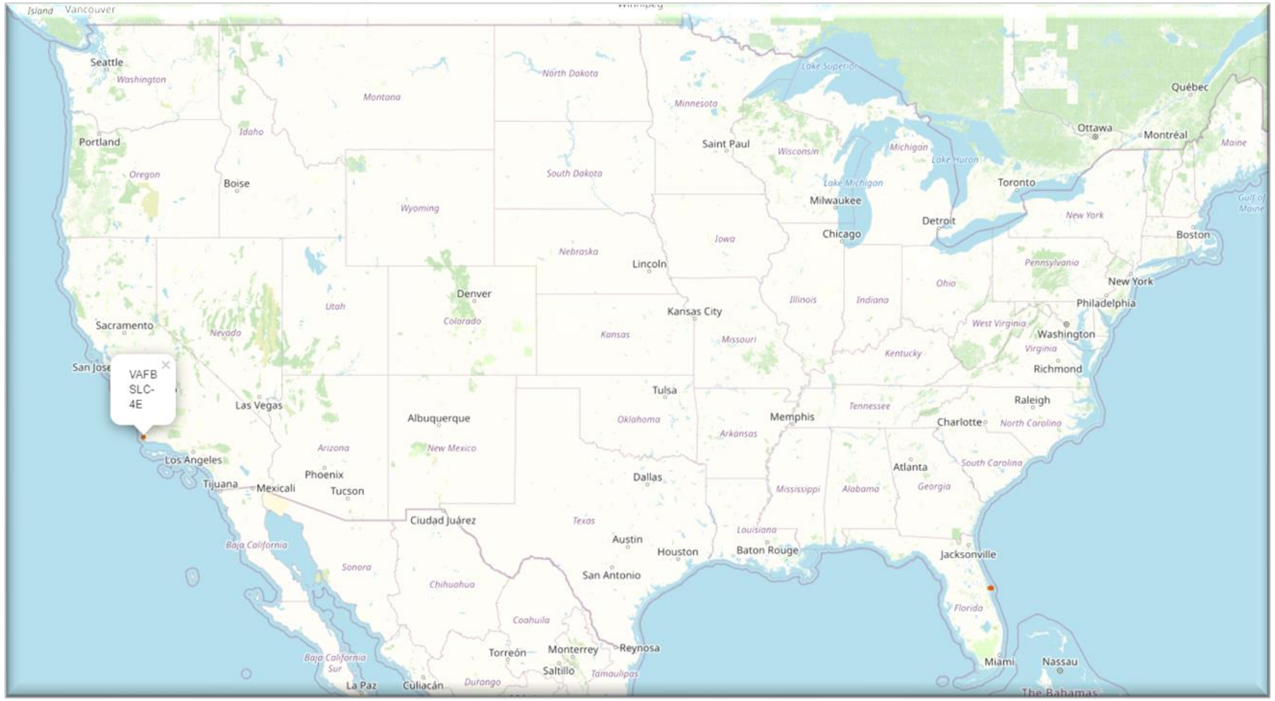
Landing _Outcome	COUNT(1)
Success	20
Success (drone ship)	8
Success (ground pad)	6

Section 3

Launch Sites Proximities Analysis

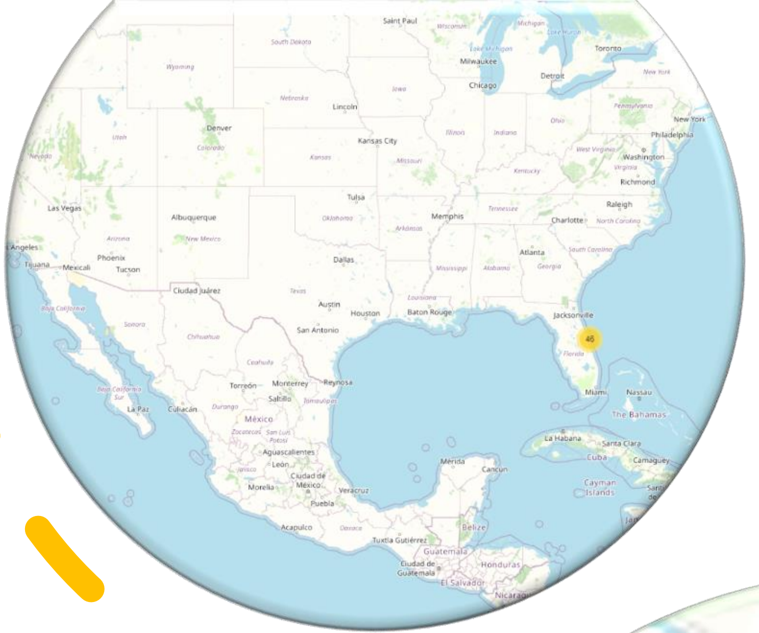
ALL LAUNCH SITES ON A MAP

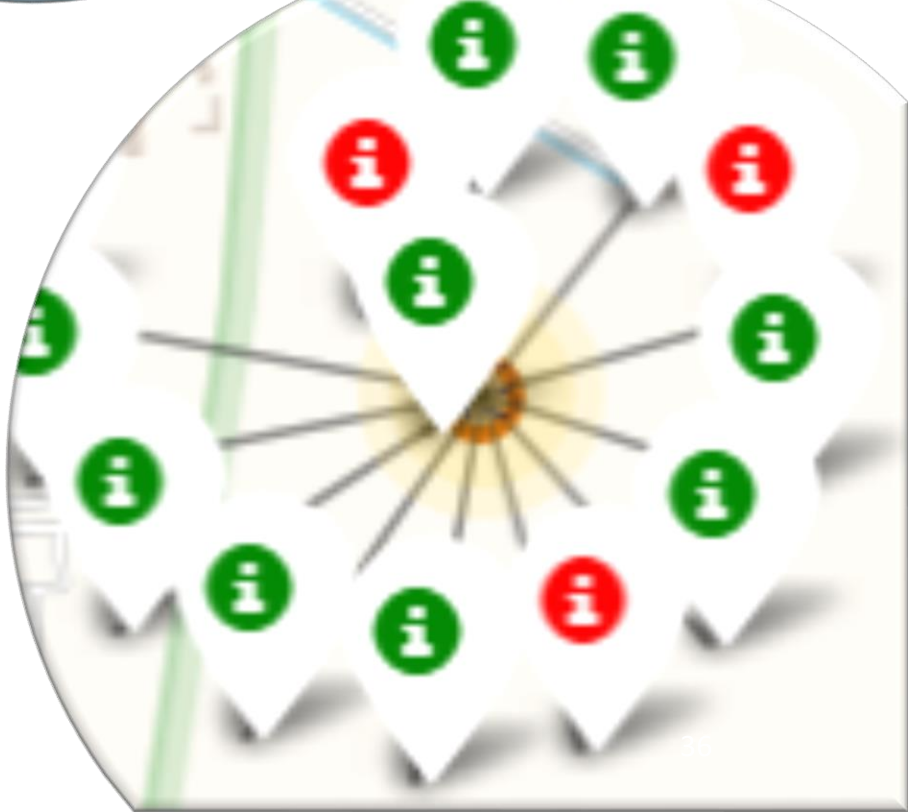
All SpaceX launch sites are on coasts of the United States of America, specifically Florida and California.



SUCCESS/FAILED LAUNCHES FOR EACH SITE

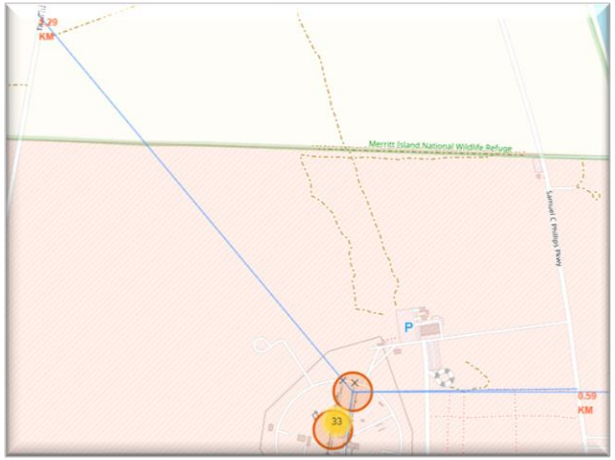
Launches have been grouped into clusters, and annotated with green icons for successful launches, and red icons for failed launches

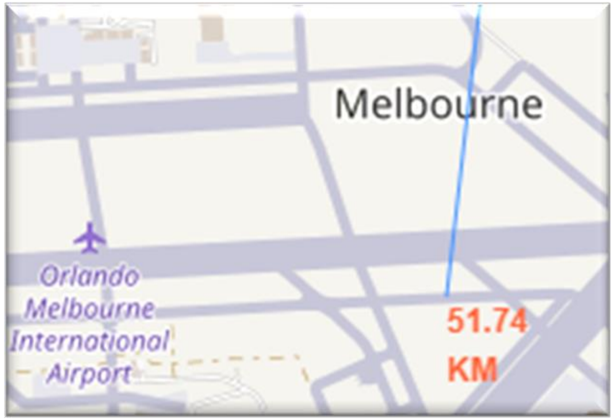


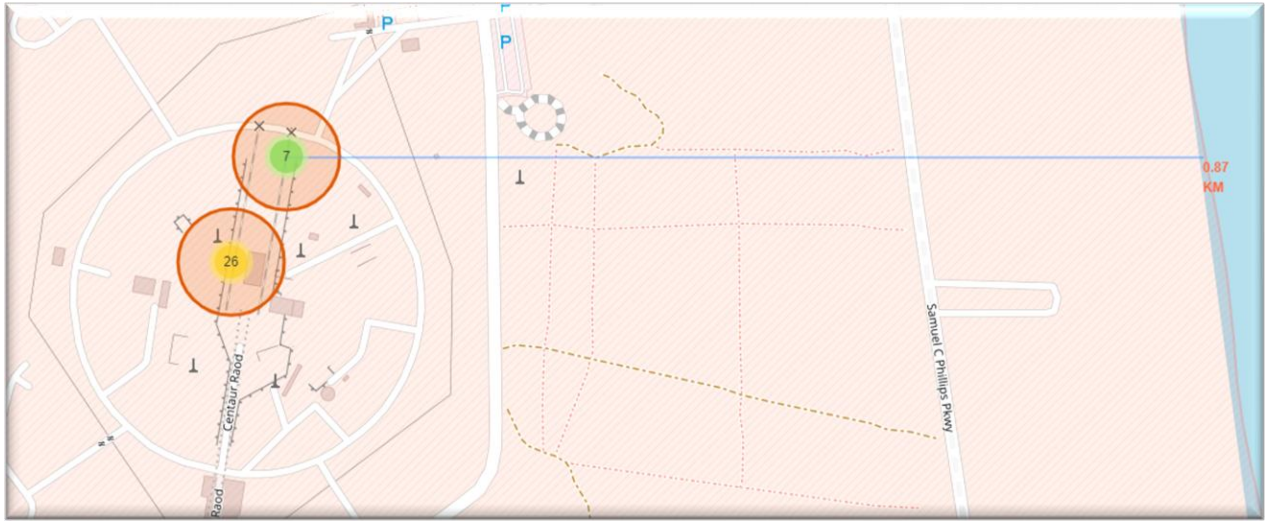


PROXIMITY OF LAUNCH SITES TO OTHER POINTS OF INTEREST

Using the CCAFS SLC-40 launch site as an example site, we can understand more about the placement of launch sites.

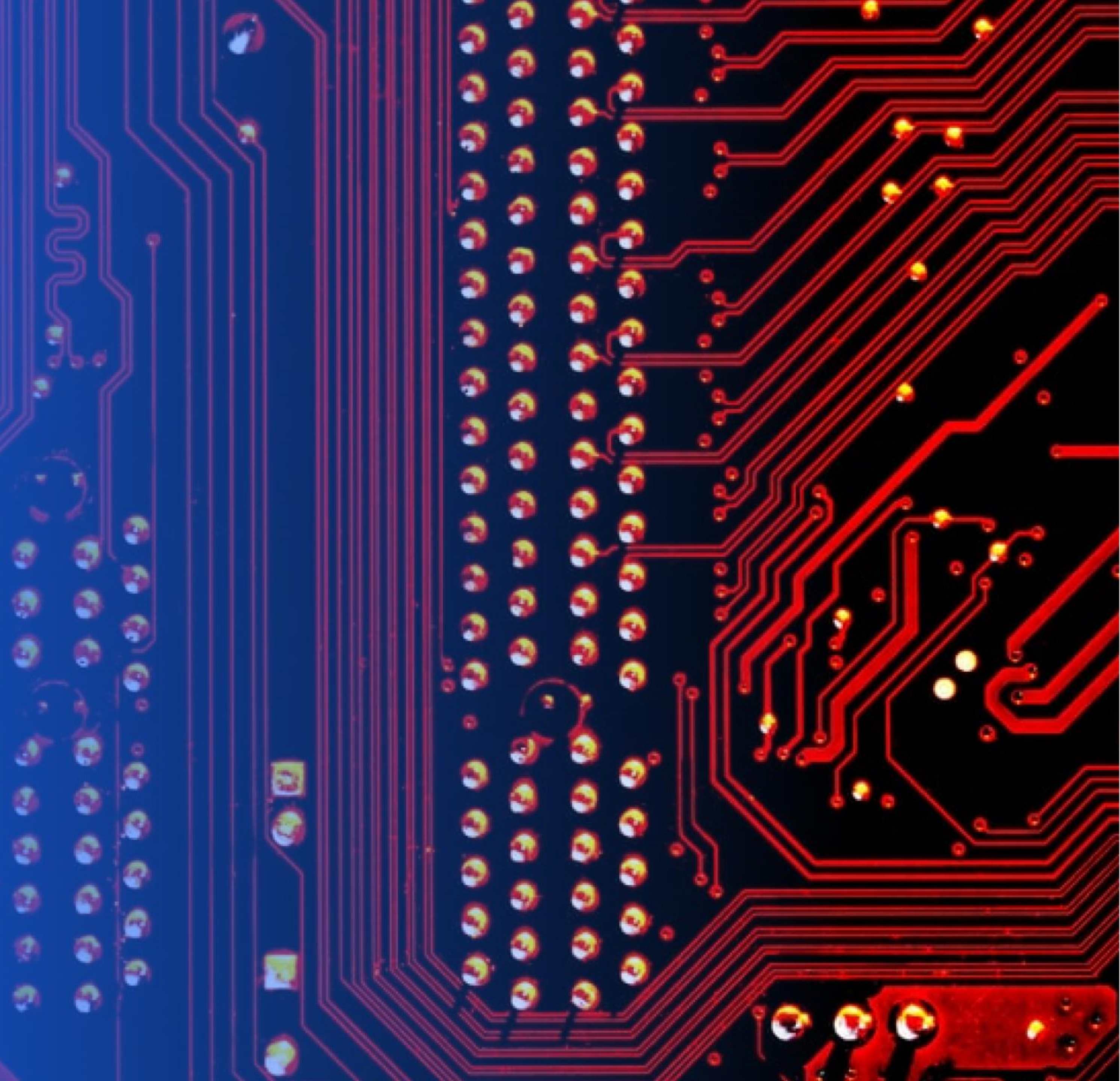




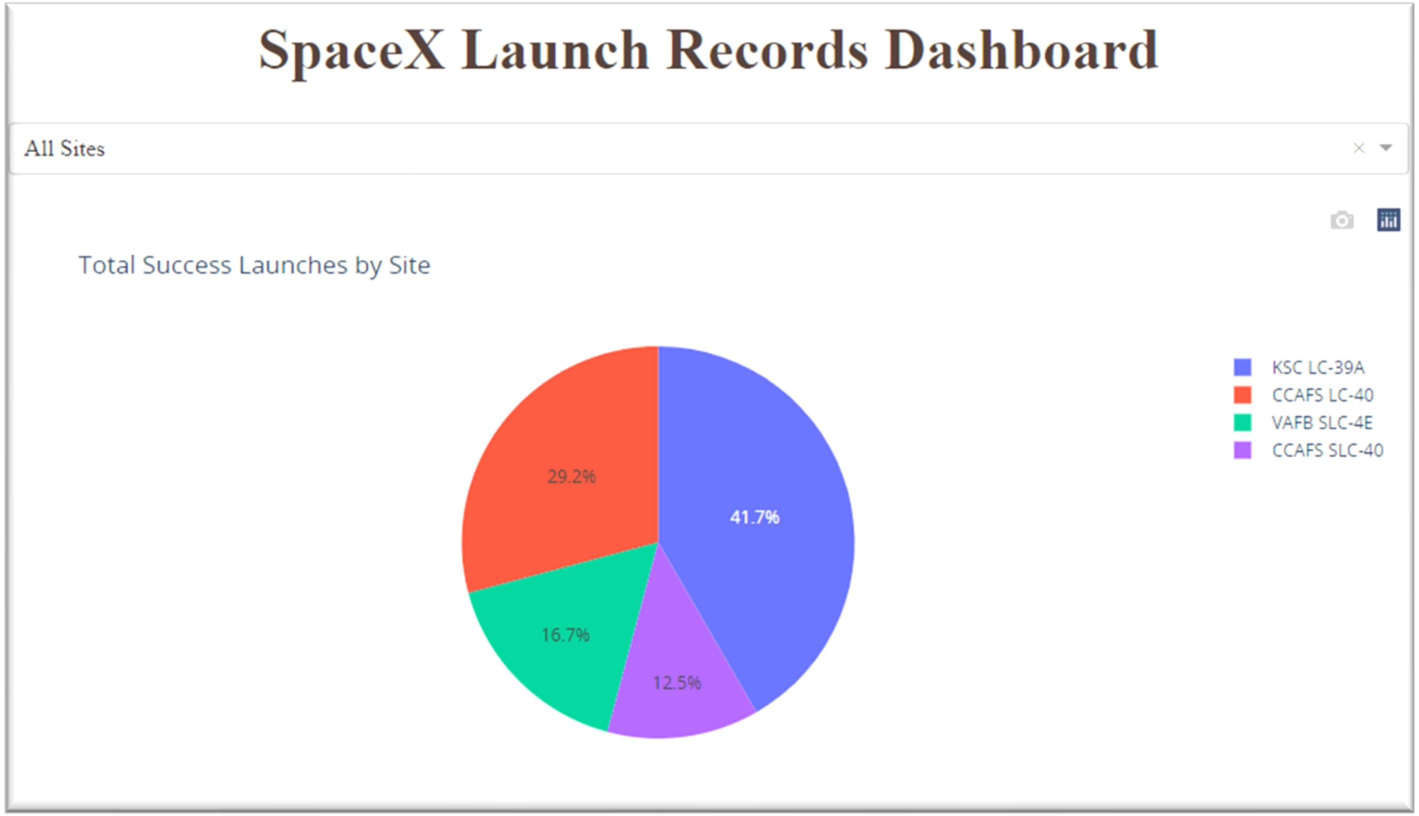


Section 4

Build a Dashboard with Plotly Dash



launch success count for all sites

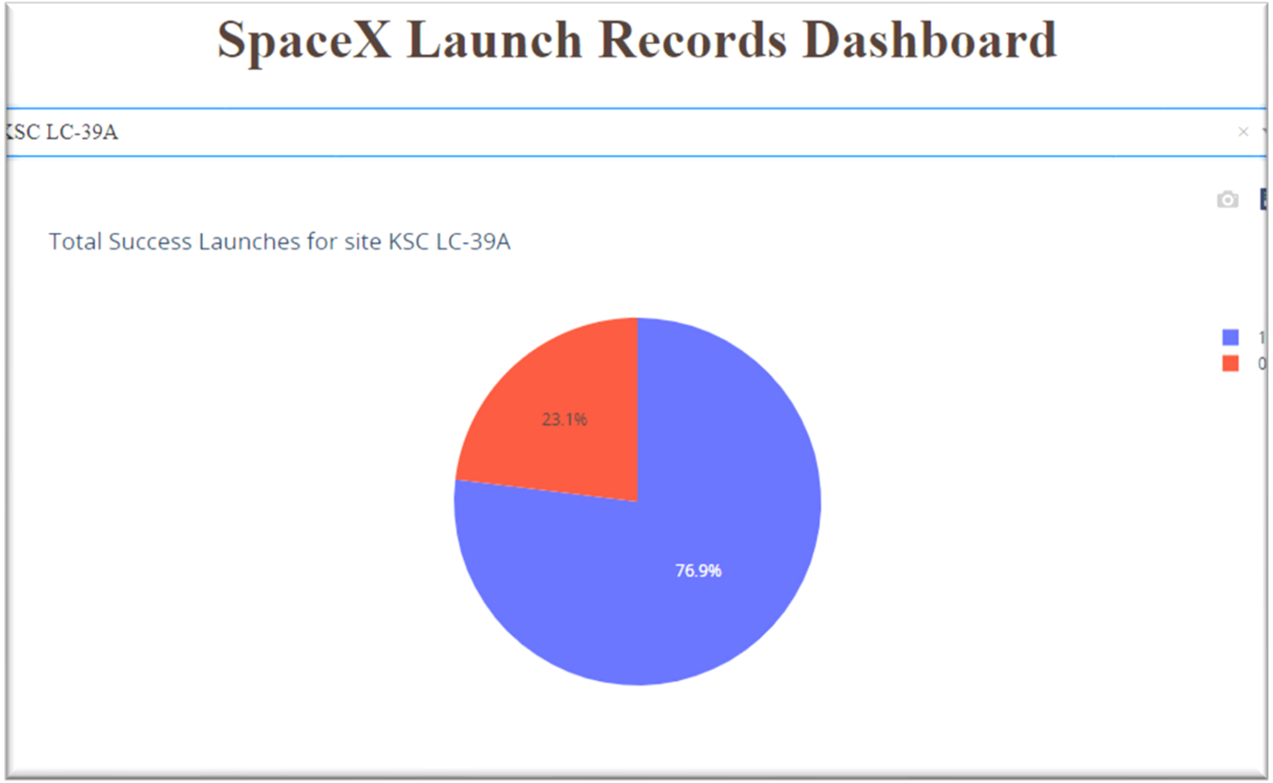


SpaceX Launch Records Dashboard

All Sites

The launch site KSC LC-39 A had the most successful launches, with 41.7% of the total successful launches.

Pie chart for the launch site with highest launch success ratio



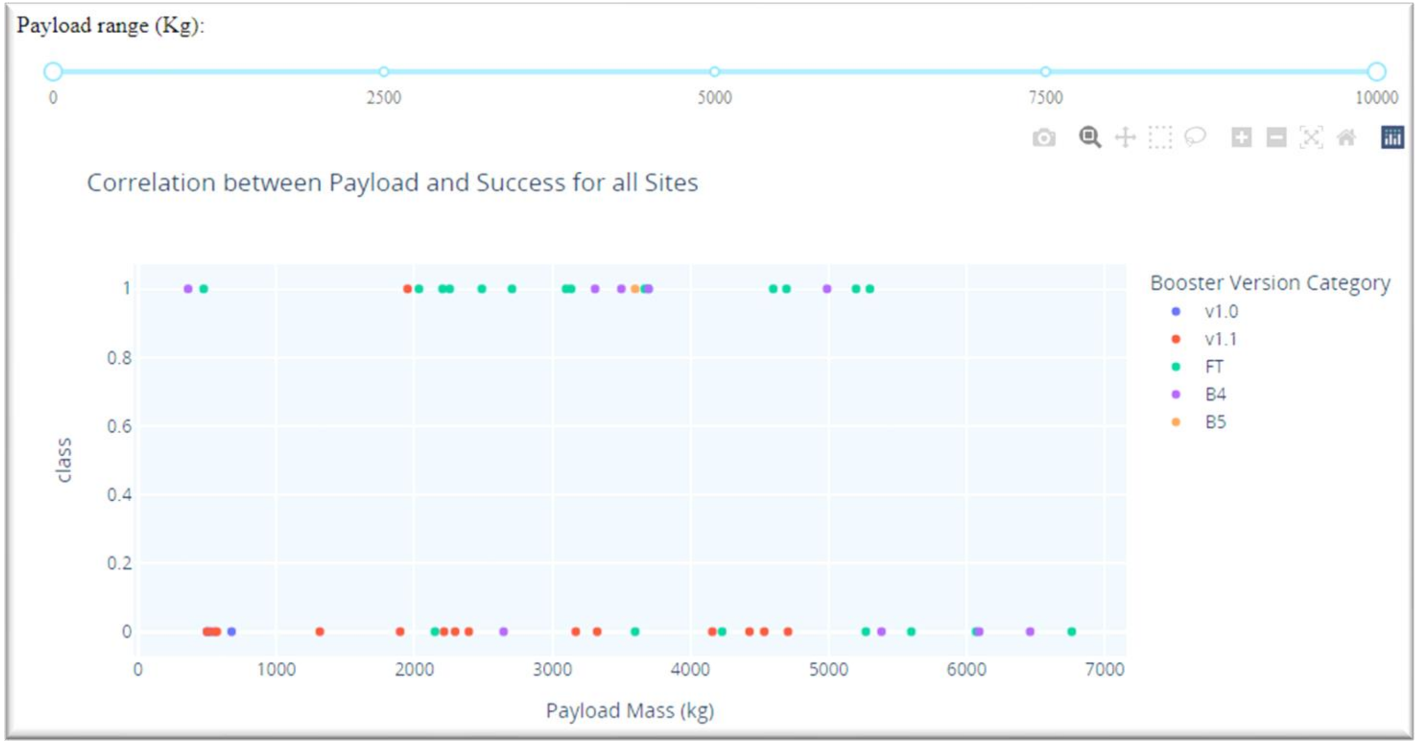
SpaceX Launch Records Dashboard

ΚSC LC-39A

Total Success Launches for site KSC LC-39A

The launch site KSC LC-39 A also had the highest rate of successful launches, with a 76.9% success rate.

Launch Outcome VS. Payload scatter plot for all sites

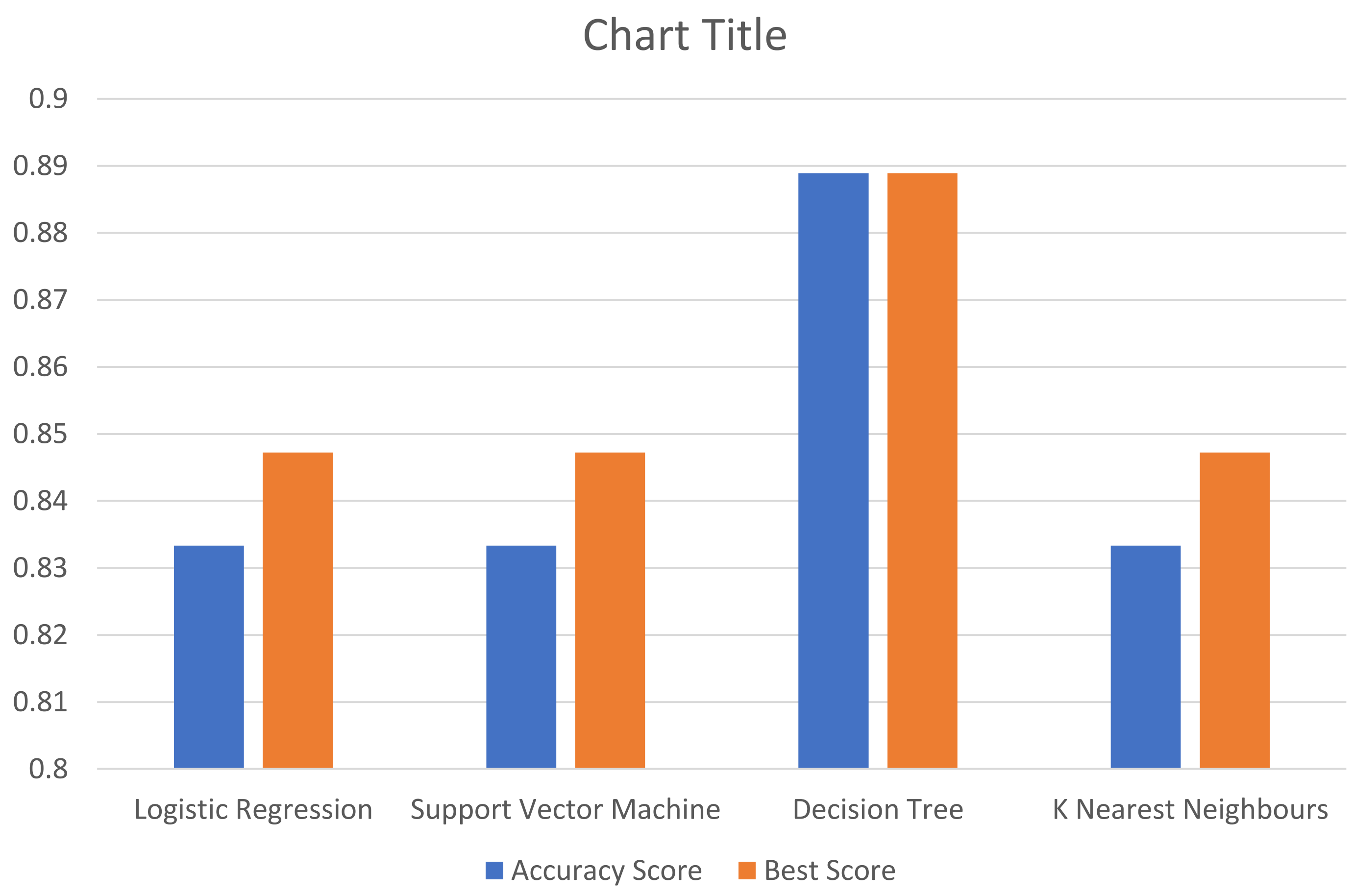


- Plotting the launch outcome vs. payload for all sites shows a gap around 4000 kg, so it makes sense to split the data into 2 ranges:
  - 0 – 4000 kg (low payloads)
  - 4000 – 10000 kg (massive payloads)
- From these 2 plots, it can be shown that the success for massive payloads is lower than that for low payloads.
- It is also worth noting that some booster types (v1.0 and B5) have not been launched with massive payloads.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

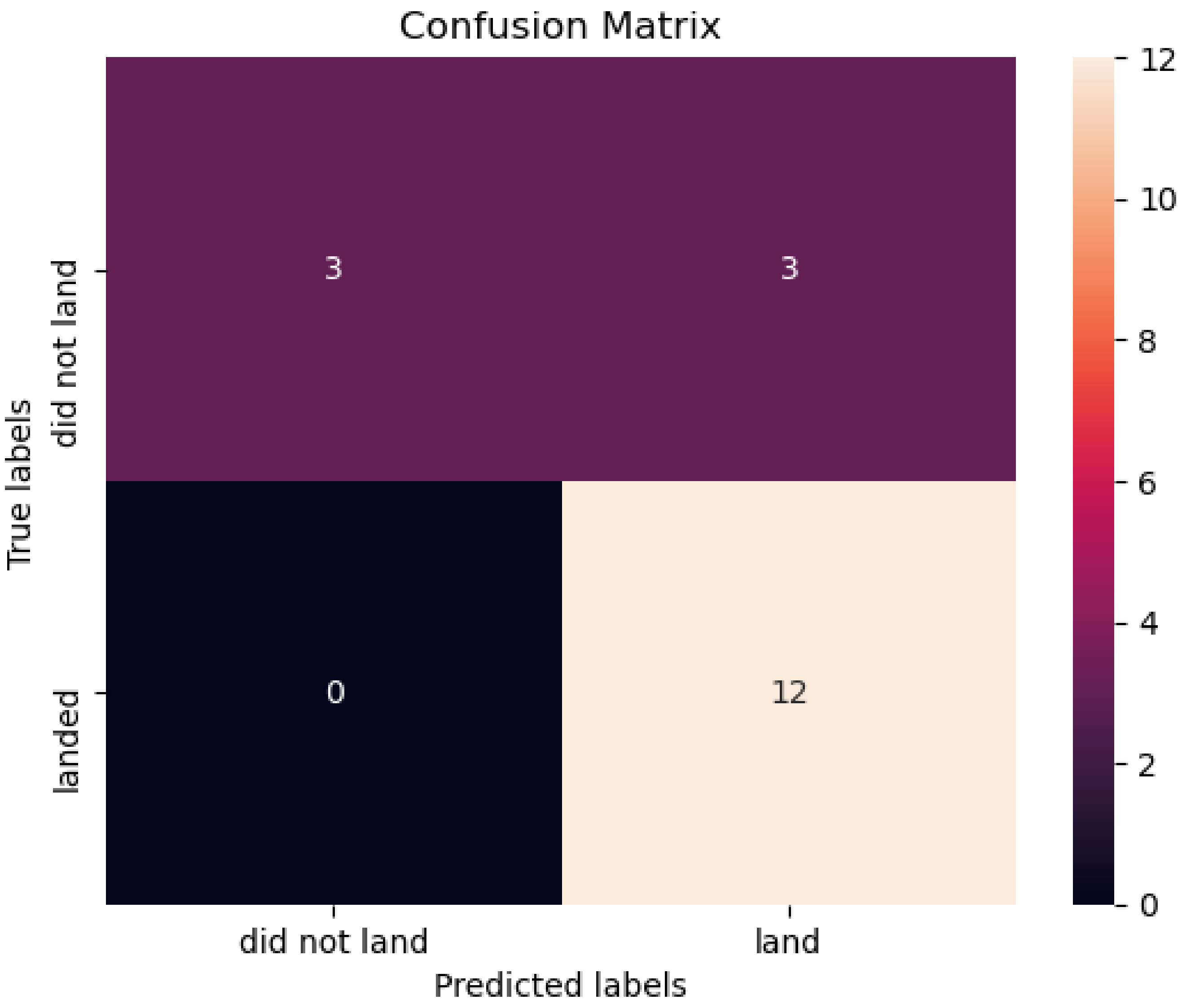


Plotting the Accuracy Score and Best Score for each classification algorithm produces the following result:

- The Decision Tree model has the highest classification accuracy
  - The Accuracy Score is 0.89
  - The Best Score is 0.89

Confusion Matrix

- As shown previously, best performing classification model is the Decision Tree model, with an accuracy score of 0.89
- This is explained by the confusion matrix, which shows only 3 out of 18 total results classified incorrectly (a false positive, shown in the top-right corner).
- The other 15 results are correctly classified (3 did not land, 12 did land).



Conclusions

- As the number of flights increases, the rate of success at a launch site increases, with most early flights being unsuccessful. I.e. with more experience, the success rate increases.
  - Between 2010 and 2013, all landings were unsuccessful (as the success rate is 0).
  - After 2013, the success rate generally increased, despite small dips in 2018 and 2020.
  - After 2016, there was always a greater than 50% chance of success.

- Orbit types ES-L1, GEO, HEO, and SSO, have the highest (100%) success rate.
  - The 100% success rate of GEO, HEO, and ES-L1 orbits can be explained by only having 1 flight into the respective orbits.
  - The 100% success rate in SSO is more impressive, with 5 successful flights.
  - The orbit types PO, ISS, and LEO, have more success with heavy payloads:
  - VLEO (Very Low Earth Orbit) launches are associated with heavier payloads, which makes intuitive sense.

- The launch site KSC LC-39 A had the most successful launches, with 41.7% of the total successful launches, and also the highest rate of successful launches, with a 76.9% success rate.

- The success for massive payloads (over 4000kg) is lower than that for low payloads.

- The best performing classification model is the Decision Tree model, with an accuracy score of 0.89.

Thank you!